INFORMATION AND CONTROL 36, 119–132 (1978)

Combinatorial Multiple-Valued Filing Systems for Multiattribute Queries

GERALD BERMAN

Department of Combinatorics and Optimization, University of Waterloo, Waterloo, Ontario, Canada N2L 3GI

Combinatorial filing systems for records with unequal numbers of attribute values are constructed in terms of cyclic groups. The notion of group-invariant sets of queries is defined, which includes sets of queries of fixed order. The buckets which are constructed from a unique decomposition of a group-invariant set of queries, and the corresponding subbuckets, can be represented in a simple way in terms of the parameters of the group representation. This leads to effective storage and retrieval algorithms which do not require scanning lists of buckets or records. The total number of buckets required and the redundancy of a cyclic filing system are usually less than those of the corresponding systems constructed in terms of combinatorial configurations. Cyclic intersection systems are constructed to handle the case in which the set of queries is not group invariant.

INTRODUCTION

A *filing system* has five components, a set of records R which represents the information or data, a set of queries Q which represents the set of retrievals required, a set of buckets B which represents the (computer) organization of the data, a storage algorithm AS for determining where the accession numbers of the records are to be stored in the buckets, and a retrieval algorithm AR for determining where the accession numbers of the records satisfying a given query are to be found.

In a *multiple-valued filing system* each record $r \in R$ is characterized by a vector $(r_1, r_2, \ldots, r_m)$ whose components provide information about the m attributes $\alpha_1, \alpha_2, \ldots, \alpha_m$. The ith *attribute* α_i is assumed to have n_i values $a_{ij}, j = 1, 2, \ldots, n_i$, and for each i, $r_i = a_{ij}$ for some j. A *query* $q \in Q$ is a subset of $A = \{a_{ij}\}$ such that no two elements of q are levels of the same attribute. The *order* of q is $| q |$, the cardinality of q.

The accession numbers of the records are stored in *buckets* $B_1, B_2, \ldots, B_b$ which are partitioned into *subbuckets*. The following conditions are usually assumed.

(1) The accession number of a record occurs at most once in a bucket;

0019-9958/78/0362-0119$02.00/0

(2) If $q \in Q$, every record $r \in R(q)$ satisfying this query occurs in the same bucket $B(q)$;

(3) If $r \in R(q) \cap P$, where P is a subbucket of $B(q)$, then every record in P (as represented by its accession number) satisfies q.

The triple (A, Q, B) which characterizes the combinatorial aspects of the system F will be referred to as a *combinatorial filing scheme* and we shall sometimes write $F = F(A, Q)$. Different algorithms AS, AR may be used in implementing the filing scheme as a filing system for arbitrary sets of records with attribute set A. A *combinatorial filing system* is a filing system based on a combinatorial filing scheme. The term combinatorial is used for the filing system as defined by its parameters to distinguish it from a computerized implementation of the system.

The storage algorithm AS is the rule for determining for each $r \in R$ the subset $J(r)$ of $I(b) = \{1, 2, ..., b\}$ which specifies the buckets B_j, $j \in J(r)$, and locates the subbuckets into which (the accession number of) r is to be placed. The retrieval algorithm AR is the rule for determing $B(q)$ for each $q \in Q$ and the collection of subbuckets containing (the accession numbers of) records satisfying q.

The simplest type of bucket structure occurs in the *first-order inverted filing scheme* obtained by letting a bucket correspond to each element of A. A record is stored in each of the buckets associated with its attribute levels $r_1, r_2, ..., r_m$. Such systems are efficient for queries specified in terms of one attribute value but are very inefficient for handling queries involving several attributes, especially for large systems. This problem can be partially solved by employing a *Q-inverted filing scheme* $F_I = F_I(A, Q)$ obtained by letting a bucket correspond to each element of Q. This clearly makes the retrieval very simple, but usually requires too many buckets.

The problem of specifying buckets to satisfy conditions (1) and (2) above is a problem of combinatorics: Given m sets $\alpha_1, \alpha_2, ..., \alpha_m$ of sizes $n_1, n_2, ..., n_m$ and a collection Q of subsets of $\alpha_1 \cup \alpha_2 \cup \cdots \cup \alpha_m$, how can b groups $B_1, B_2, ..., B_b$ be formed such that any subset $q \in Q$ will be contained in one and only one of the b groups, say $B(q)$. Further, it is required that there exist a "simple" algorithm for determining $B(q)$ for every $q \in Q$.

If $Q = Q^{(k)}$ is the set of all queries of order k, then $F = F(A, Q^{(k)})$ is a *balanced multiple-valued filing* (BMVF) *scheme* of order k. Various special methods have been published for constructing suitable sets B. Bose *et al.* (1967) employed equations over the finite field $GF(n)$ to construct systems with equal parameters $n_1 = n_2 = \cdots = n_m = n$. Ghosh and Abraham (1968) employed finite geometries to develop BMVF schemes of order 2, also with equal parameters, and Abraham *et al.* (1968) generalized to schemes of order k using linear forms over $GF(n)$. Ghosh (1969) used deleted finite geometries to construct systems with unequal parameters.

Another approach was initiated by Ray-Chaudhuri (1968). He employed combinatorial configurations for the case of binary valued attributes. Bose and Koch (1969) extended this idea to handle multiple-valued attributes. Yamamoto *et al.* (1972) used cyclically generated spreads in finite projective geometries to construct BMVF systems of order 2. The storage and retrieval algorithms for these systems were simplified by Berman (1976a). Analogous systems were constructed by Berman (1976b) using PBIB designs.

In this paper Q is not restricted to $Q^{(k)}$. Combinatorial filing systems are constructed for any set of parameters $n_1, n_2, \ldots, n_m$ in terms of cyclic groups. The group representation of A is given in Section 2 and the corresponding properties of Q are considered in Section 3. The notion of a group-invariant set of queries is defined, and a unique decomposition contructed into subsets whose elements are cyclically ordered. In Section 4 buckets and subbuckets are constructed from the decomposition of a group-invariant set of queries. The storage and retrieval algorithms are formulated in Section 5 in terms of parameters of the group representation. The algorithms do not require scanning the buckets or lists of records. The definition of group invariance includes the case of sets of queries of order k, so that as a special case there are cyclic BMVF systems of order k.

Redundancy is discussed in Section 6 and approximate formulas are obtained for cyclic BMVF systems of order k with equal parameters which are uniformly distributed in R. Comparisons are made with corresponding systems constructed in terms of combinatorial configurations. In Section 7 it is shown how to handle systems in which Q is not group invariant as cyclic intersection systems. This approach can also be used to simplify systems with unequal parameters.

2. Group Representation of A, R

Renumber the attributes $\alpha_0, \alpha_1, \ldots, \alpha_{m-1}$ so that

$$\alpha_i = (a_{i1}, a_{i2}, \ldots, a_{in_i}), \qquad i = 0, 1, \ldots, m-1, \tag{2.1}$$

and

$$n = n_0 \geqslant n_1 \geqslant \cdots \geqslant n_{m-1}. \tag{2.2}$$

Let m_j denote the number of nonzero values in the sequence $a_{0j}, a_{1j}, a_{2j}, \ldots$. Then

$$m = m_1 \geqslant m_2 \geqslant \cdots \geqslant m_n. \tag{2.3}$$

Let

$$C_j = \{u_j, u = 0, 1, \ldots, m_j - 1\} \tag{2.4}$$

denote the elements of the cyclic group G_j of order m_j, $j = 1, 2,..., n$; that is, u_j is the residue class of elements congruent to u modulo m_j. Set

$$a_{ij} = i_j, \qquad j = 1, 2,..., n_i; \qquad i = 0, 1,..., m - 1 \tag{2.5}$$

(or $i = 0, 1,..., m_j - 1$ for $j = 1, 2,..., n$). By (2.1) and (2.4) the attributes then have the representation

$$\alpha_u = (u_1, u_2, ..., u_{n_u}), \qquad u = 0, 1,..., m - 1, \tag{2.6}$$

where $u_j \in C_j$, $j = 1, 2,..., n_u$, and A is given by

$$A = \begin{bmatrix} 0_1 & 0_2 & \cdots & 0_{n_0} \\ 1_1 & 1_2 & \cdots & 1_{n_1} \\ \vdots & \vdots & \vdots & \vdots \\ (m-1)_1 & (m-1)_2 & \cdots & (m-1)_{n_{m-1}} \end{bmatrix} \tag{2.7}$$

or

$$A = \begin{bmatrix} 0_1 & 0_2 & \cdots & 0_n \\ 1_1 & 1_2 & \cdots & 1_n \\ \vdots & \vdots & \vdots & \vdots \\ (m_1-1)_1 & (m_2-1)_2 & \cdots & (m_n-1)_n \end{bmatrix}. \tag{2.8}$$

This is the *group representation* of the attributes. The entries of column j of (2.8) (or (2.7)) are the elements of C_j, $j = 1, 2,..., n$.

If the parameter values are equal, i.e., $n_1 = n_2 = \cdots = n_m = n$, then $m_1 = m_2 = \cdots = m_n = m$ and $G_1 = G_2 = \cdots = G_n$ is the cyclic group of order m. In this case we shall write $A = A(m, n)$ so that

$$A(m, n) = \{u_j\}, \qquad u = 0, 1,..., m - 1; j = 1, 2,..., m. \tag{2.9}$$

The records $t \in R$ have the (group) representation

$$r = (0_{j_0}, 1_{j_1}, ..., (m-1)_{j_{m-1}}), \qquad 1 \leqslant j_i \leqslant n_i, \quad i = 0, 1,..., m - 1, \tag{2.10}$$

or

$$r = \{u_j \in C_j, u = 0, 1,..., m - 1\}. \tag{2.11}$$

That is, r contains one entry from each row of A in (2.7) (or (2.8)).

EXAMPLE 2.1. Consider a set of records with five attributes having 6, 4, 6, 4, and 3 values. These can be renumbered $\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4$ with $n_0 = n_1 = 6$, $n_2 = n_3 = 4$, $n_4 = 3$ to satisfy (2.2). Then $m = 5$ and $m_1 = m_2 = m_3 = 5$,

$m_4 = 4$, $m_5 = m_6 = 2$ satisfy (2.3). Taking cyclic groups of order m_j, $j = 1, 2,..., 6$, the attribute values have representation (2.7) or (2.8), which is given by

$$A = \begin{bmatrix} 0_1 & 0_2 & 0_3 & 0_4 & 0_5 & 0_6 \\ 1_1 & 1_2 & 1_3 & 1_4 & 1_5 & 1_6 \\ 2_1 & 2_2 & 2_3 & 2_4 & & \\ 3_1 & 3_2 & 3_3 & 3_4 & & \\ 4_1 & 4_2 & 4_3 & & & \end{bmatrix}, \tag{2.12}$$

where $u_j \in G_j$, $j = 1, 2,..., 6$. Each record can be represented as $r = (a_0, a_1, a_2, a_3, a_4)$, where a_i is an element of the ith row of (2.12), $i = 0, 1, 2, 3, 4$.

3. Group-Invariant Sets of Queries

Let Q be a set of queries and χ the mapping from A to A defined by

$$\chi u_j = (u + 1)_j, \qquad u_j \in C_j. \tag{3.1}$$

Set

$$\chi q = \{\chi x, x \in q\}, \qquad q \in Q. \tag{3.2}$$

χ is the *cyclic operator*. The set of queries Q is *group invariant* if $\chi q \in Q$ for every $q \in Q$. In particular the set $Q^{(k)}$ of all queries of order k is group invariant.

Let Q be group invariant. There is a unique subset $Q(q) \subset Q$ for every $q \in Q$, defined by

$$Q(q) = \{\chi^j q, j = 0, 1,...\}. \tag{3.3}$$

The cardinality of $Q(q)$, the *period* $\gamma(q)$ of q relative to χ, depends on the period of the elements of q as well as the distribution of the elements in the cyclic groups. The following theorem is immediate.

Theorem 3.1. *Let* $q \in Q$ *and* $p_j = q \cap C_j$, $j = 1, 2,..., n$, *so that* $q = p_1 \cup p_2 \cup \cdots \cup p_n$. *Then*

$$\gamma(q) = \text{LCM}(k_1, k_2,..., k_n),$$

where k_j *is the period of* p_j, *relative to* χ, *i.e.,* k_j *is the smallest positive integer such that* $\chi^{k_j} p_j$, $j = 1, 2,..., n$. *Further, if* $| q | = k$ *and every element of* q *has the same period* t *such that* $\text{GCD}(k!, t) = 1$, *then* $\gamma(q) = t$.

Example 3.2. Consider a set of records as in Example 2.1 and the query $q = \{2_1, 4_2, 1_4, 3_4, 0_6\}$. In this case $k_1 = k_2 = 5$, $k_4 = k_6 = 2$, and $k_3 = k_5 = 1$

($p_3 = p_5 = \varnothing$), so that $\gamma(q) = 10$. There are queries of order 2 with periods 2, 4, 5, 10, and 20; for example, $\{0_5, 0_6\}$, $\{0_4, 0_5\}$, $\{0_1, 0_2\}$, $\{0_3, 0_5\}$, $\{0_3, 0_4\}$. The queries $\{0_1, 1_1\}$, $\{0_1, 0_2\}$ which have period 5 illustrate the second part of the theorem. Notice also that although $k = 2$, $t = 4$ for $(0_4, 1_5)$ the GCD$(2,4) = 2$ and the period is 4, but in the case $(0_4, 2_4)$ the period is 2.

The cyclic operator defines an equivalence relation which partitions Q uniquely into disjoint subsets. Take $q_1 \sim q_2$ if $q_1 \in Q(q_2)$. It is immediate from (3.3) that $q_2 \in Q(q_1)$ and $Q(q_1) = Q(q_2)$. The corresponding partition

$$Q = Q_1 \cup Q_2 \cup \cdots \cup Q_b \tag{3.4}$$

is the *cyclic partition* of Q. Select a representative query q_s from each set Q_s, $s \in I(b)$. For example, let q_s be the first element of Q_s after the elements have been arranged in some lexicographic order. Then every element $q \in Q$ has a unique representation in the form

$$q = \chi^t q_s, \qquad 0 \leqslant t \leqslant \gamma_s - 1, s \in I(b), \tag{3.5}$$

where $\gamma_s = \gamma(q_s)$ is the period of q_s relative to χ. Set $\gamma_0 = 0$.

Let σ, τ denote the mappings defined by (3.5) and

$$\sigma(q) = s, \qquad \tau(q) = t. \tag{3.6}$$

The pair σ, τ determines a 1–1 mapping ν from Q to the set of integers $I(|Q|)$ defined by

$$\nu(q) = \sum_{j<\sigma(q)} \gamma_j + \tau(q) + 1. \tag{3.7}$$

The function ν has a unique inverse ν^{-1} from $I(|Q|)$ to Q.

The above remarks are summarized in Theorem 3.3.

THEOREM 3.3. *Let Q be a group-invariant set of queries. The cyclic operator χ determines a unique partition of Q into disjoint sets* (3.4). *Every element $q \in Q$ has a unique representation in the form* (3.5) *for any set of distinct representatives $\{q_j\}$ of $\{Q_j\}$. This determines two mappings σ, τ from Q to I given by* (3.6) *and a* 1–1 *mapping ν from Q onto $I(|Q|)$ given by* (3.7).

The values of $\sigma(q)$, $\tau(q)$ can be expressed in terms of the parameters $\gamma_1, \gamma_2, \ldots, \gamma_b$. Set

$$\delta_1 = 0, \qquad \delta_j = \sum_{i<j} \gamma_i, \qquad j = 2, 3, \ldots, b+1. \tag{3.8}$$

Then

$$0 = \delta_1 < \delta_2 < \cdots < \delta_b < \delta_{b+1} = |Q|. \tag{3.9}$$

THEOREM 3.4. *Let the queries be represented by the integers* $q^* \in I(|Q|)$, *with* $\nu(q) = q^*$. *Then by* (3.9) *there is an integer* s *such that*

$$\delta_s < q^* \leqslant \delta_{s+1}, \qquad 1 \leqslant s \leqslant b, \tag{3.10}$$

so that $q^* = \delta_s + t + 1$ *and* $\sigma(q) = s$, $\tau(q) = t$.

EXAMPLE 3.5. Suppose $A = A(11, 8)$. By (2.9) and (2.7) or (2.8), A is given by

$$A = \begin{bmatrix} 0_1 & 0_2 & \cdots & 0_8 \\ 1_1 & 1_2 & \cdots & 1_8 \\ \cdot & \cdot & \cdots & \cdot \\ 10_1 & 10_2 & \cdots & 10_8 \end{bmatrix}, \tag{3.11}$$

where G_j is isomorphic to the cyclic group of 11 elements for every $j \in I(8)$. Let $Q = Q^{(2)}$ so that the 3520 queries have the representation $\{u_i, v_j\}$, $u \neq v$, $i, j \in I(8)$. It is easy to verify that the cyclic operator partitions $Q^{(2)}$ into 320 disjoint sets each containing 11 queries which may be represented in the form B_{ij}^u ; $i, j \in I(8)$, $u \in I(5)$. A convenient representative of the set Q_{ij}^u is q_{ij}^u, given by

$$q_{ij}^u = \{0_i, u_j\}, \qquad i, j \in I(8), \quad u \in I(5). \tag{3.12}$$

These 320 queries can be relabeled $q_1, q_2, \ldots, q_{320}$ and the corresponding subsets of $Q^{(2)}$ can be relabeled $Q_1, Q_2, \ldots, Q_{320}$ by setting

$$q_s = q_{ij}^u, \qquad s = i + 8(j-1) + 64(u-1), \qquad i, j \in I(8), \quad u \in I(5). \tag{3.13}$$

In this case $\nu_1 = \nu_2 = \cdots = \nu_{320} = 11$ and $\delta_j = 11(j-1)$, $j \in I(321)$, and the queries are labeled 1, 2,..., 3520.

To illustrate the mappings σ, τ, ν consider the query $q = \{7_2, 4_5\}$. This may be rewritten as

$$q = \chi^4\{0_5, 3_2\} = \chi^4 q_{5,2}^3 = \chi^4 q_{141},$$

using (3.12) and (3.13). Thus $\sigma(q) = 141$, $\tau(q) = 4$ and by (3.7) $\nu(q) = 140 \cdot 11 + 4 + 1 = 1545$, i.e., $q^* = 1545$. To illustrate the inverse mapping ν^{-1}, let $q = 1124$. Applying Theorem 3.4 we find that 1545 lies between $\delta_{141} = 1540$ and $\delta_{142} = 1551$. Thus $1545 = \delta_{141} + 5$, implying $\sigma(q) = 141$, $\tau(q) = 4$. In base 8 the integer 141 has the representation $141 = 5 + 1 \cdot 8 + 2 \cdot 8^2$ so that by (3.13), $i = 5$, $j = 2$, $u = 3$, and $q_{141} = q_{5,2}^3$. By (3.12), $q_{5,2}^3 = \{0_5, 3_2\}$, implying $q = \chi^4\{0_5, 3_2\} = \{4_5, 7_2\}$.

The above example illustrates a general class of systems (BMVF of order 2) which have convenient representations. The following two theorems summarize the general idea involved, and will be used to construct cyclic BMVF schemes.

THEOREM 3.6. *Let $A = A(m, n)$ be the group representation of m attributes each with n values and let $Q = Q^{(k)}(m, n)$ be the corresponding set of queries of order k. Then Q is group invariant. Further, if* $\mathrm{GCD}(k!, m) = 1$ *then every query has period m and $\gamma_s = m$, $s = 1, 2, \ldots, b$, where*

$$b = \frac{n^k}{m}\binom{m}{k}. \tag{3.14}$$

This is an immediate consequence of Theorem 3.1.

THEOREM 3.7. *Let $A = A(m, n)$, $Q = Q^{(2)}(m, n)$ and suppose m is odd. Then the cyclic decomposition of Q has $n^2(m-1)/2$ sets which can be labeled Q_{ij}^u, $i, j \in I(n)$, $u \in I((m-1)/2)$, and a set of representatives is given by*

$$q_{ij}^u = \{0_i, u_j\}, \qquad i, j \in I(n), \quad u \in I\left(\frac{m-1}{2}\right). \tag{3.15}$$

These queries may be relabeled q_s, $s \in I(n^2(m-1)/2)$, where

$$q_s = q_{ij}^u, \qquad s = i + (j-1)n + (u-1)n^2 \tag{3.16}$$

for $i, j \in I(n)$, $u \in I((m-1)/2)$. Further, $\gamma_1 = \gamma_2 = \cdots = \gamma_b = m$, so that if the queries $q \in Q$ are labeled $1, 2, \ldots, |Q|$, the mappings σ, τ, ν are related by

$$q = \nu(q) = (\sigma(q) - 1)m + \tau(q) + 1, \qquad |\tau(q)| < m. \tag{3.17}$$

In particular, $\tau(q)$ is the remainder when q is divided by m.

Theorem 3.7 can easily be extended to even values of m, say $m = 2m'$. Formula (3.15) holds for $u \in I(m'-1)$. To obtain a unique representation for $u = m'$ the condition $i \leqslant j$ is required. The queries $q_{jj}^{m'}$, $j \in I(n)$, have period m' and all the others have period m as before. Because of this, formulas (3.16) and (3.17) require modification.

Explicit formulas for σ, τ, ν can be given in certain special cases.

THEOREM 3.8. *Let $A = A(m, n)$, $Q = Q^{(2)}(m, n)$, m odd. Let $q = \{u_i, v_j\}$, $u < v$, denote any query. Then the values of $\sigma(q)$, $\tau(q)$, $\nu(q)$* are given by

$$\sigma(q) = k + (l-1)n + (y-1)n^2, \tag{3.18}$$

$$\tau(q) = x, \tag{3.19}$$

$$\nu(q) = [(k-1) + (l-1)n + (y-1)n^2]m + x + 1, \tag{3.20}$$

where

$$x = u, \quad y = v - u, \quad k = i, \quad l = j \qquad \text{for} \quad v - u \leqslant \frac{m-1}{2}$$

and

$$x = v, \quad y = m + u - v, \quad k = j, \quad l = i \qquad \text{for} \quad v - u \geqslant \frac{m-1}{2}.$$

This follows ar once from Theorems 3.3 and 3.7 and the observation that $\{u_i, v_j\} = \chi^u\{0_i, (v-u)_j\}$ if $v - u \leqslant (m-1)/2$ and $\{u_i, v_j\} = \{v_j, (m+u)_i\} = \chi^v\{0_j, (m+u-v)_i\}$ with $m + u - v \leqslant (m-1)/2$ if $v - u \geqslant (m-1)/2$.

4. Buckets and Subbuckets

Let A be a group representation of the attributes and let Q be a group-invariant set of queries with cyclic partition (3.4). With each set Q_j associate a bucket B_j, $j = 1, 2, \ldots, b$, of B. A record $r \in R$ will be stored in every bucket B_s such that r satisfies a query $q \in Q_s$. With each subset $P \subset Q_s$ associate a subbucket $B_s(P)$ of B_s. The record r is stored in $B_s(P)$ if P is the subset of all queries of B_s satisfied by r. To indicate the dependence on A, Q the set B will also be denoted by $B(A, Q)$, or $B(m, n)$, $B^{(k)}(m, n)$ in the special cases in which $A = A(m, n)$ and $A = A^{(k)}(m, n)$, respectively. The corresponding *cyclic filing scheme* is denoted by (A, Q, B) or $F_c(A, Q)$, and in the special cases $F_c(m, n)$, $F_c^{(k)}(m, n)$.

It is convenient to put a label $\lambda_s(P)$ on the subbucket $B_s(P)$ for easy identification. By Theorem 3.3 every element $p \in Q_s$ has a unique representation in the form

$$p = \chi^{\tau(p)} q_s, \qquad 0 \leqslant \tau(p) < \nu_s, p \in Q_s. \tag{4.1}$$

Set

$$\lambda_s(P) = \sum_p 2^{\tau(p)}, \qquad p \in P \subset Q_s. \tag{4.2}$$

These labels are the integers $1, 2, \ldots, 2^{|Q_s|} - 1$.

Bucket Structure 4.1. The buckets $B_1, B_2, \ldots, B_b$ are in 1–1 correspondence with the sets $Q_1, Q_2, \ldots, Q_b$ of the cyclic decomposition of Q. The subbuckets of B_s are in 1–1 correspondence with the subsets of Q_s, the subset $P \subset Q_s$ having the label $\lambda_s(P)$.

Example 4.2. In the filing scheme $F^{(2)}(5, 3)$, Q contains 90 queries which can be labeled $1, 2, \ldots, 90$ and 18 buckets $B_1, B_2, \ldots, B_{18}$. By Theorem 3.7, ν can be chosen so that $\nu(q) = q$ and σ, τ satisfy the equation

$$q = 5\sigma(q) + \tau(q) - 4, \qquad |\tau(q)| < 5, \tag{4.3}$$

and $Q_s = \{5s - j, 0 \leqslant j \leqslant 4\}$, so that $\sigma(5s - j) = s$, $\tau(5s - j) = 4 - j$,

$j = 0, 1, 2, 3, 4$. Consider the query $\{0_2, 1_2\}$. As in Example 3.5, $\{0_2, 1_2\} = \chi^0 q_{22}^1 = \chi^0 q_s$ so that $\sigma\{0_2, 1_2\} = 5$ and $\{0_2, 1_2\} \in Q_5$. The queries $\{0_2, 4_2\}$ and $\{1_2, 2_2\}$ are also in Q_5. The values of τ for these three queries are 0, 4, 1 so that the subset $P \subset Q_5$ containing these queries has label $\lambda_5(P) = 2^0 + 2^4 + 2^1 = 19$.

5. Storage and Retrieval

Let (A, Q, B) be a cyclic filing scheme as described in Section 4. Let R denote any set of records with attribute values in A so that $r \in R$ is given by (2.10) or (2.11), and let σ, τ, ν denote the mappings from Q to I as defined in Section 4.

Storage Algorithm 5.1. Let $r \in R$ denote any record.

AS1 Determine the set $Q(r) \subset Q$ of all queries $q \in Q$ such that r satisfies the query q.

AS2 Compute $\sigma(q)$, $\tau(q)$ for each $q \in Q(r)$ and partition $Q(r)$ into subsets $P_s(r) \subset Q_s$, $S \in I(b)$ so that

$$Q(r) = P_1(r) \cup P_2(r) \cup \cdots \cup P_b(r).$$

AS3 For each s such that $P_s(r) \neq \varnothing$ store the accession number of r in the subbucket of B_s with label $\lambda_s(P_s(r))$ (given by 4.2 in terms of σ, τ).

Example 5.2. Consider the system $F_c^{(2)}(5, 3)$ and suppose R contains the record $r = (0_2, 1_2, 2_2, 3_3, 4_2)$. Then $Q(r)$ consists of the 10 queries $\{0_2, 1_2\}$, $\{0_2, 2_2\}$,..., $\{3_3, 4_2\}$ which we shall denote by $P_1, P_2, \ldots, P_{10}$. Using the method illustrated in Example 4.2 it is easy to verify that these queries are in the sets $Q_5, Q_{14}, Q_{15}, Q_5, Q_5, Q_{17}, Q_{14}, Q_8, Q_{14}, Q_6$, respectively. Thus $P_6(r) = \{P_{10}\}$, $P_5(r) = \{P_1, P_4, P_5\}$, $P_8(r) = \{P_8\}$, $P_{14}(r) = \{P_2, P_7, P_9\}$, $P_{15}(r) = \{P_3\}$, $P_{17}(r) = \{6\}$, and the remaining $P_j(r) = \varnothing$. In Example 4.2 it was shown that $\lambda(P_5(r)) = 19$. In the same way it can be shown that $\lambda(P_6(r)) = 8$, $\lambda(P_8(r)) = 4$, $\lambda(P_{14}(r)) = 21$, $\lambda(P_{15}(r)) = 8$, and $\lambda(P_{17}(r)) = 2$. Thus the accession number of r is stored in the buckets $B_6, B_5, B_8, B_{14}, B_{15}$, and B_{17} in the subbuckets 8, 19, 4, 21, 8, and 2 respectively.

In applying AS, Storage Algorithm 5.1, the buckets are computed directly from r (employing the functions σ, τ) without checking the buckets. This property is also true of the following retrieval algorithm.

Retrieval Algorithm 5.3. Let $q \in Q$.

AR1 Determine $s = \sigma(q)$, $t = \tau(q)$, and γ_s the order of Q_s.

AR2 Calculate the set of integers

$$S(q) = \Big\{2^t + \sum_i \epsilon_i 2^i,\ \epsilon_i = 0, 1;\ 0 \leqslant i < \gamma_s,\ i \neq t\Big\}.$$

AR3 Retrieve all accession numbers in all subbuckets of B_s with label in $S(q)$.

The set $S(q)$ contains the labels of all subsets of B_s containing q, i.e., the labels of all subbuckets of B_s containing q. These are the subbuckets into which the accession numbers of records corresponding to the query q were stored by AS.

As illustrated in Example 5.2 the storage algorithm can be stated more explicitly for cyclic BMVF systems of order 2 by making use of the representation theorems 3.7 and 3.8. In this case $A = A(m, n)$, $Q = Q^{(2)}(m, n)$, $B = B^{(2)}(m, n)$, and $r = (r_1, r_2, \ldots, r_m)$. The set $Q(r)$ of AS1 is then given by $Q(r) = \{q = \{r_i, r_j\}, i \neq j\}$. The representation of Q as described in Theorems 3.7 and 3.8 leads to simplification of AS2 and AR1.

6. Redundancy

The *redundancy* $\rho = \rho(F)$ of a filing system is the average number of buckets into which the records are placed, i.e.,

$$\rho(F) = \sum_r J(r)/|\ R\ |, \qquad r \in R, \tag{6.1}$$

where $J(r) \subset I(b)$ is the number of buckets into which r is placed. This may also be written as

$$\rho(F) = \sum_s K_s/|\ R\ |, \qquad s \in I(b), \tag{6.2}$$

where K_s is the number of records of R stored in B_s. For a given set R, $\rho(F)$ depends on Q as well as B. The value of $\rho_I = \rho(F_I)$ (F_I is the Q-inverted filing scheme) provides an upper bound for $\rho(F)$ for a given pair (R, Q), for if more than one query corresponds to a bucket of F, e.g., $B(q_1) = B(q_2)$, any record which corresponds to both q_1 and q_2 will be stored once in F but at least twice in F_I.

The redundancy also depends on the distribution of the attribute values in the records and is difficult to calculate for general R, Q. For this reason it is usually assumed that the attribute values are uniformly distributed in R and $Q = Q^{(k)}$. Then $\rho_I = \binom{m}{k}$ provides an upper bound to $\rho(F)$ for any filing system based on $(R, Q^{(k)})$.

Rather than consider a fixed set of records R, another approach expresses ρ in terms of the probability that a record will correspond to a given query.

Setting $p_s = K_s/|R|$, (6.2) can be rewritten as

$$\rho(F) = \sum_s p_s, \qquad s \in I(b), \tag{6.3}$$

and p_s can be interpreted as the probability that a record will be stored in B_s. The sum in (6.3) is then the expected times the accession number of the record is stored in the buckets.

An approximation to (6.3) for F_c can be obtained as follows. If $q \in Q_s$ has order k, and if r is any record (given by (2.10)) then r satisfies q if the subscripts in q and r coincide. Assuming a uniform distribution of records, the probability of this is n^{-k}. Since there are γ_s queries in Q_s, the probability that r satisfies at least one of the queries in Q_s is approximately given by $p_s = 1 - (1 - n^{-k})^{\gamma_s}$. This expression is exact for $k = 1$ and is a good approximation for $k > 1$, especially if m is large and the k-tuples in r are almost independent. Thus

$$\rho(F_c) \doteq \sum_s (1 - (1 - n^{-k})^{\gamma_s}), \qquad s \in I(b). \tag{6.4}$$

This simplifies for the case $F = F^{(k)}(m, n)$ if $\mathrm{GCD}(k!, m) = 1$, for by Theorem 3.6, $\gamma_s = m$ and b is given by (3.14). Then

$$\rho^{(k)}(m, n) \doteq n^k \binom{m}{k} (1 - (1 - n^{-k})^m)/m, \tag{6.5}$$

where $\rho^{(k)}(m, n) = \rho(F^{(k)}(m, n))$. In particular, for $k = 1, 2$,

$$\rho^{(1)}(m, n) = n(1 - (1 - n^{-1})^m), \tag{6.6}$$

$$\rho^{(2)}(m, n) \doteq n^2(m - 1)(1 - (1 - n^{-2})^m)/2. \tag{6.7}$$

EXAMPLE 6.1. Consider the system $F^{(2)}(5, 3)$. In this case $\rho^{(2)}(5, 3) = 8.0$, by (6.7). To compute the exact value we may proceed as follows. Consider the bucket Q_{11}^1 defined by q_{11}^1 containing the queries $\{u_1, (u + 1)_1\}$, $u = 0, 1, 2, 3, 4$ (mod 5). The total number of different records in R is $3^5 = 243$ (assuming one record of each type in a uniform distribution). The records corresponding to $\{0_1, 1_1\}$ have the form $(0_1, 1_1, 2_i, 3_j, 4_k)$. There are exactly 12 which do not contain another query in Q_{11}^1, namely the records for which $i = 1, 2$; $j = 1, 2, 3$; $k = 1, 2$. By analogy there are also 12 records having exactly one query in common with Q_{11}^1 for each of the other queries $\{u_i, (u + 1)_i\}$, $i = 1, 2, 3, 4$. Thus there are 60 records having exactly one query in common with Q_{11}^1. In the same way it can be verified that there are 20 records having two pairs in common with Q_{11}^1, 10 records with three pairs, no records with four pairs, and 1 record with five pairs, showing that the number of records stored in B_{11}^1 is $K_{11}^1 = 91$. The number of pairs involved is $L_{11}^1 = 60 + 2(20) + 3(10) + 5(1) = 135$. For the bucket B_{12}^1 there are one or two possibilities, one or two queries in common

with a record; 21 records have one given query in common with Q_{12}^1 and 3 records have two given pairs in common with Q_{12}^1, so that $K_{12}^1 = 120, L_{12}^1 = 135$. All 6 buckets B_{ii}^k have the same L_{ii}^k, K_{ii}^k as B_{11}^1 and all 12 buckets B_{ij}^k, $i \neq j$, have the same values of K_{ij}^k, L_{ij}^k as Q_{12}^1. This accounts for all queries since $\sum_{i,j} L_{ij}^k = 2430 = 243 \times 10$. By (6.2), $\rho^{(2)}(5, 3) = (6K_{11}^1 + 12K_{12}^1)/243 = 8.2$. This compares with a value of 10.0 for $\rho_I^{(2)}(5, 3)$.

To compare with the system F_s of order 2 based on cyclically generated spreads of finite projective geometry as described in Yamamoto *et al.* (1972) and Berman (1976a), let $Q^{(12)}(m, n) = Q^{(1)}(m, n) \cup Q^{(2)}(m, n)$. Table I shows the values of b_c, ρ_c for $F_c(A, Q^{(12)}(m, n))$ (to the nearest integer) together with the corresponding values of ρ_s, b_s, ρ_I, b_I for all odd values of $m < 100$ for which the system based on spreads exists.

TABLE I

Numbers of Buckets and Redundancy for $F_c^{(12)}(m, n), F_s^{(12)}(m, n), F_I^{(12)}(m, n)$

m	n	b_c	b_s	b_I	ρ_c	ρ_s	ρ_I
5	3	21	30	105	11	10	15
9	7	203	588	1827	38	39	45
17	5	205	340	3485	105	96	153
17	15	1815	10200	30855	142	143	153
21	3	93	630	1953	85	173	231
33	31	15407	169136	508431	540	548	561
65	21	14133	91728	918645	1956	1942	2145
73	7	1771	42924	129283	1379	2431	2701
85	3	381	10710	32385	381	2814	3655
91	4	724	10920	65884	722	2896	4186

7. Cyclic Intersection Systems

If $\hat{Q}$ is a set of queries which is not group invariant, set

$$Q = \{x^j q, q \in \hat{Q}, j = 0, 1, \ldots\}.$$

Then Q is group invariant. Let (3.4) be the cyclic decomposition of Q and let

$$\hat{Q}_s = Q_s \cap \hat{Q}, \qquad s \in I(b),$$

so that $\hat{Q} = \hat{Q}_1 \cup \hat{Q}_2 \cup \cdots \cup \hat{Q}_s$. Let $\hat{B}$ denote the corresponding buckets. The

combinatorial scheme $\hat{F}_c = (A, \hat{Q}, \hat{B})$ is the *cyclic intersection system* associated with $\hat{Q}$. Since $\hat{F}_c$ coincides with F_c in $\hat{Q}$ it follows that $\hat{F}_c$ may be represented by the parameters of F_c allowing for void sets in Q and the buckets and subbuckets of B.

The representation of a cyclic system as an intersection system of a simpler cyclic system (e.g., $F^{(2)}(m, n)$, m odd, as defined by Theorem 3.7) may also be convenient in cases of unequal parameters. Notice that any value of n may be used by replacing attributes α_i with n_i values $n_i > n$ by the union of attributes with smaller numbers of values.

EXAMPLE 7.1. Consider a system with five attributes as in Example 2.1 for which $n_0 = n_1 = 6$, $n_2 = n_3 = 4$, $n_4 = 3$. This can be represented by the system $F_c(7, 4)$ by setting $\alpha_0 = (\beta_0, \beta_5)$, $\alpha_1 = (\beta_1, \beta_6)$ where β_0, β_1 have four values and β_5, β_6 two values. Another representation is by the system $F_c(5, 6)$, where additional values (not used) are added to the attributes $\alpha_2, \alpha_3, \alpha_4$. In the case $Q = Q^{(2)}$ the buckets of $F_c^{(2)}(7, 4)$ or $F_c^{(2)}(5, 6)$ can be conveniently represented as in Theorem 3.7.

RECEIVED: September 3, 1976; REVISED: March 18, 1977

REFERENCES

ABRAHAM, C. T., GHOSH, S. P., AND RAY-CHAUDHURI, D. K. (1968), File organization schemes based on finite geometries, *Inform. Contr.* **12**, 143–163.

BERMAN, G. (1976*a*), The application of difference sets to the design of a balanced multiple-valued filing scheme, *Inform. Contr.* **32**, 128–138.

BERMAN, G. (1976*b*), "Group Generated Balanced Combinatorial Configurations and Information Retrieval Systems for Files," Research Report CORR 76/10, University of Waterloo.

BOSE, R. C., ABRAHAM, C. T., AND GHOSH, S. P. (1967), File organization of records with multiple valued attributes for multi-attribute queries, *in* "Proceedings of a Symposium on Combinatorial Mathematics," pp. 277–297, Univ. of North Carolina Press, Chapel Hill.

BOSE, R. C., AND KOCH, G. G. (1969), The design of combinatorial information retrieval systems for files with multiple valued attributes, *SIAM J. Appl. Math.* **17**, 1203–1214.

GHOSH, S. P. (1969), Organization of records with unequal multiple-valued attributes and combinatorial queries of order 2, *Inform. Sci.* **1**, 363–380.

GHOSH, S. P., AND ABRAHAM, C. T. (1968), Application of finite geometry in file organization for records with multiple-valued attributes, *IBM J. Res. Develop.* **12**, 180–187.

RAY-CHAUDHURI, D. K. (1968), Combinatorial information retrieval systems for files, *SIAM J. Appl. Math.* **16**, 973–992.

YAMAMOTO, S., TERAMOTO, T., AND FUTAGAMI, K. (1972), Design of a balanced multiple-valued filing scheme or oder 2 based on cyclically generated spread in finite projective geometry, *Inform. Contr.* **21**, 72–91.